Toric Ideals of Homogeneous Phylogenetic Models

Nicholas Eriksson
Department of Mathematics
University of California, Berkeley
Berkeley, CA 94720-3840
eriksson@math.berkeley.edu

ABSTRACT

We consider the model of phylogenetic trees in which every node of the tree is an observed, binary random variable and the transition probabilities are given by the same matrix on each edge of the tree. The ideal of invariants of this model is a toric ideal in $\mathbb{C}[p_{i_1 \ldots i_n}]$. We are able to compute the Gröbner basis and minimal generating set for this ideal for trees with up to 11 nodes. These are the first non-trivial Gröbner bases calculations in $2^{11} = 2048$ indeterminates. We conjecture that there is a quadratic Gröbner basis for binary trees, but that generators of degree n are required for some trees with n nodes. The polytopes associated with these toric ideals display interesting finiteness properties. We describe the polytope for an infinite family of binary trees and conjecture (based on extensive computations) that there is a universal bound on the number of vertices of the polytope of a binary tree.

1. INTRODUCTION

A phylogenetic tree is a rooted tree T on n nodes with a κ-ary random variable X_i associated to every node. Write $\rho(v)$ for the parent of node v. Then the transition probabilities between $\rho(v)$ and v are given by a κ by κ matrix $A^{(v)}$ for every non-root node of T.

In an application, κ might encode the four nucleic acids that make up DNA, the two families of nucleic acids, or the twenty amino acids. The transition matrices are generally picked from some specific family such as the Jukes-Cantor [9], Kimura [10], or general Markov models [1].

In this paper we consider the homogeneous Markov model where all $A^{(v)}$ are equal, all nodes are binary ($\kappa = 2$) and observable, and the root has uniform distribution. We write $A^{(v)} = A = \begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix}$. The probability of observing i at a node v is computed from the parent of v by

$$P(X_v = i) = a_{0i} P(X_{\rho(v)} = 0) + a_{1i} P(X_{\rho(v)} = 1).$$

We are interested in the algebraic relations satisfied by the joint distribution

$$p_{i_1 i_2 \ldots i_n} := P(X_1 = i_1, \ldots, X_n = i_n).$$

Writing the joint distribution in terms of the model parameters $a_{00}, a_{01}, a_{10}, a_{11}$, we have

$$p_{i_1 i_2 \ldots i_n} = \prod_{j=2}^{n} a_{i_{\rho(j)} i_j} \tag{1}$$

where the nodes of the tree are labeled 1 to n starting with the root. That is, the probability of observing a certain labeling of the tree is the product of the a_{ij} that correspond to the transitions on all edges of the tree. The indeterminates a_{ij} parameterize a toric variety of dimension 4 in $\mathbb{R}^{2^n}$. We let I_T be the corresponding toric ideal, called the ideal of phylogenetic invariants. In the notation of [11], the toric ideal I_T is specified by the 4 by 2^n configuration $\mathcal{A}_T$, where column $(i_1, \ldots, i_n)$ consists of the exponent vector of the a_{ij} in (1). We order the rows $(a_{00}, a_{01}, a_{10}, a_{11})$. Let P_T be the convex hull of the columns of $\mathcal{A}_T$.

We are interested in two questions from [8]. First, which relations on the joint probabilities $p_{i_1 \ldots i_n}$ does the model imply? This problem is solved by giving generators of the ideal of invariants I_T.

In Section 2, we study the generators of this ideal. Our main accomplishment is the computation of Gröbner and Markov bases for trees with 11 nodes. These are computations in 2048 indeterminants, which we believe to be the largest number of indeterminants ever in a Gröbner basis calculation. We also calculate generating sets for all trees on at most 9 nodes. Based on this evidence, we conjecture that if T is binary, then the ideal I_T has a quadratic generating set, and furthermore, that relations of degree n are necessary to generate I_T for certain trees with n nodes.

Our second goal is to determine, given a labeling of the tree T, if we can identify parameters a_{ij} such that the labeling is the most likely among all labelings? This problem is solved by computing the normal fan of the toric variety in the sense of [4].

In Section 3, we study this normal fan and the polytope P_T. Our main result, Theorem 1, is an explicit description of the polytope P_T for an infinite family of binary trees. For this family, P_T always has 8 vertices and 6 facets which we characterize. We also present extensive calculations of P_T

for various trees and conjecture that there is a bound on the number of vertices of P_T as T ranges over all binary trees.

The invariants vanish for a given distribution $(p_{i_1\ldots i_n})$ essentially when that distribution comes from our model. Thus the knowledge of the generators of this ideal is potentially very useful for fitting biological sequence data to a phylogenetic tree, as first noted by Cavender and Felsenstein [2]. While there has been much progress towards finding the ideal of invariants for other phylogenetic models (see [1], [9], [10]), the homogeneous model is particularly attractive because the low number of parameters makes it possible to compute non-trivial examples. Hopefully we can use the homogeneous model to approximate in some sense the general model, perhaps by subdividing edges of the tree.

Example 1. Let T be a path with 3 nodes. Then

$$\mathcal{A}_T = \begin{pmatrix} 2 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 2 \end{pmatrix},$$

the polytope P_T has 7 vertices and 6 facets, and the toric ideal of the path of length 3 is generated by 6 binomials

$$\begin{aligned} I_T = \langle & x_{101} - x_{010}, x_{001}x_{100} - x_{000}x_{010}, \\ & x_{011}x_{100} - x_{001}x_{110}, x_{011}x_{110} - x_{010}x_{111}, \\ & x_{001}^2x_{111} - x_{000}x_{011}^2, x_{100}^2x_{111} - x_{000}x_{110}^2\rangle. \end{aligned}$$

2. TORIC IDEALS

The toric ideals I_T are homogeneous, since all monomials in (1) have the same degree $n-1$. Thus they define projective toric varieties Y_T. Algebraic geometers usually require a toric variety to be normal, but the reader should be warned that the toric varieties discussed in this paper are generally not normal.

Recall that a projective toric variety given by a configuration $\mathcal{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_k)$ is covered by the affine toric varieties given by $\mathcal{A} - \mathbf{a}_i$. An affine toric variety defined by a configuration $\mathcal{A}$ is said to be smooth if the semigroup $\mathbb{N}\mathcal{A}$ is isomorphic to $\mathbb{N}^r$ for some r [12, Lemma 2.2].

PROPOSITION 1. *The projective toric variety Y_T of a binary tree T is not smooth.*

PROOF. Recall that the columns of the configuration $\mathcal{A}_T$ are indexed by 0/1-labelings of the tree T. Look at the affine chart $I_{\mathcal{A}-\mathbf{a}_{0\ldots0}}$, where $\mathbf{a}_{0\ldots0}$ corresponds to the all zero tree. On this chart, write $\tilde{\mathbf{a}}_i = \mathbf{a}_i - \mathbf{a}_{0\ldots0}$. Let $10\ldots0$ be the tree with a 1 at the root and zeros everywhere else, $0\ldots01$ be the tree with a 1 at a single leaf and zeros everywhere else, and $0\ldots010\ldots0$ be the tree with a single 1 at the parent of a leaf and zeros elsewhere. Then since $\mathbf{a}_{0\ldots0} = (n-1,0,0,0)$, we have

$$\begin{aligned} \tilde{\mathbf{a}}_{10\ldots0} &= (n-3,0,2,0) - \mathbf{a}_{0\ldots0} = (-2,0,2,0) \\ \tilde{\mathbf{a}}_{0\ldots01} &= (n-2,1,0,0) - \mathbf{a}_{0\ldots0} = (-1,1,0,0) \\ \tilde{\mathbf{a}}_{0\ldots010\ldots0} &= (n-4,1,2,0) - \mathbf{a}_{0\ldots0} = (-3,1,2,0), \end{aligned}$$

and so we see that

$$\tilde{\mathbf{a}}_{10\ldots0} + \tilde{\mathbf{a}}_{0\ldots01} = \tilde{\mathbf{a}}_{0\ldots010\ldots0}.$$

Therefore, $\mathcal{A} - \mathbf{a}_{0\ldots0}$ is not isomorphic to $\mathbf{N}^r$ and the toric variety Y_T is not smooth. □

We are primarily interested in the generators of the ideals I_T. Knowledge of the generators would allow us to easily compute whether given data came from the homogeneous Markov model from some specific phylogenetic tree.

Using `4ti2` [5], Gröbner and Markov bases for the ideal I_T were computed for all trees with at most 9 nodes as well as selected trees with 10 and 11 nodes. This took about 6 weeks of computer time in total on a 2GHz computer. The computations in 2048 variables (trees with 11 nodes) each took as long as a week and required over 2 GB of memory.

Details about the Markov bases for all binary trees with at most 11 nodes are shown in Table 1. These computations lead us to make the following conjectures.

CONJECTURE 1. *The toric ideal corresponding to a binary tree is generated in degree 2. More generally, if every non-leaf node of the tree has the same number of children d (for $d \geq 2$), the toric ideal is generated in degree 2.*

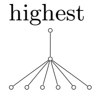
CONJECTURE 2. *There exists a quadratic Gröbner basis for the toric ideal of a binary tree.*

Using the Gröbner Walk [3] implementation in `magma`, we have computed thousands of Gröbner bases for random term orders for the smallest binary trees. It doesn't seem to be possible to compute the entire Gröbner fan for these examples with `CaTS` [6], but the random computations have yielded some information: Conjecture 2 is tree for the binary tree with 5 nodes, in fact, there are at least 4 distinct quadratic Gröbner bases for this tree. Analysis of these bases lends some optimism towards Conjecture 2. However, for the binary trees on 7 nodes, computation of over 1000 Gröbner bases did not find a quadratic basis. The best basis found contained quartics and some bases even contained relations of degree 29.

Another nice family of toric ideals is given by I_T for T a path of length n. Table 2 presents data for Markov bases of paths that leads us to conjecture that this family also has well behaved ideals.

CONJECTURE 3. *The toric ideal corresponding to a path is generated in degree 3, with $2n-4$ generators of degree 3 needed.*

Unfortunately, the toric ideal of a general tree doesn't seem to have such simple structure. For $n \leq 9$, the trees with highest degree minimal generators are those of the form



. These trees require generators of degree n.

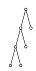
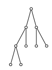
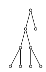
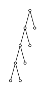
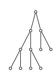
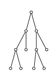
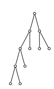
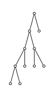
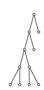
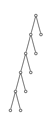
tree	Degree of I_T	#Minimal Generators	Max degree of generator
	4	4	2
	28	79	2
	92	441	2
	96	561	2
	210	2141	2
	220	2068	2
	210	2266	2
	412	7121	2
	404	7131	2
	400	7137	2
	412	7551	2
	412	7551	2
	404	7561	2

Table 1: Degree of I_T, number of minimal generators, and maximum degree of the generators

# of nodes	Degree of I_T	#Minimal Generators	Max degree	Number of deg 3
3	6	6	3	2
4	19	32	3	4
5	36	102	3	6
6	61	259	3	8
7	90	540	3	10
8	127	1041	3	12
9	168	1842	3	14
10	217	3170	3	16
11	270	5286	3	18

Table 2: Degree of I_T, size of Markov basis, maximum degree of a minimal generator, and number of degree 3 generators for paths

3. POLYTOPES

In this section, we are interested in the following problem. Given any observation $(i_1, \ldots, i_n)$ of the tree, which matrices $A = (a_{ij})$ make $p_{i_1 \ldots i_n}$ maximal among the coordinates of the distribution p?

To solve this problem, transform to logarithmic coordinates $b_{ij} = \log(a_{ij})$. Then the condition that $p_{i_1 \ldots i_n} > p_{l_1 \ldots l_n}$ for all $(l_1, \ldots l_n) \in \{0,1\}^n$ is translated into the the linear system of inequalities

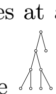
$$b_{i_1 i_2} + \cdots + b_{i_{\rho(n)} i_n} > b_{l_1 l_2} + \cdots + b_{l_{\rho(n)} l_n}$$

for all $(l_1, \ldots l_n) \in \{0,1\}^n$. The set of solutions to these inequalities is a polyhedral cone. For most values of $i_1, \ldots, i_n$, this cone will be empty. Those sequences $i_1, \ldots, i_n$ for which the cone is maximal are called *Viterbi* sequences. The collection of the cones, as $(i_1, \ldots, i_n)$ varies, is the normal fan of the polytope P_T, where P_T is the convex hull of the columns of $\mathcal{A}_T$.

Notice that P_T is a polytope in $\mathbb{R}^4$. However, since all the monomials in (1) are of degree $n-1$, we see that this polytope is actually contained in $n-1$ times the unit simplex in $\mathbb{R}^4$. Thus, P_T is actually a 3 dimensional polytope. We call P_T the *Viterbi* polytope.

The polytopes P_T show remarkable finiteness properties as T varies. Since P_T is defined as the convex hull of 2^n vectors, it would seem that it could have arbitrarily bad structure. However, as it is contained in $n-1$ times the unit simplex, it can be shown that there are at most $O(n^{1.5})$ integral points in P_T.

Example 2. Eric Kuo has shown [7] that if T is a path with n nodes, then P_T has only two combinatorial types for $n > 3$, depending only on the parity of n. The polytope for the path with 7 nodes is shown in Figure 1. Think of this picture as roughly a tetrahedron with the vertex corresponding to all $0 \to 1$ transitions and the vertex with all $1 \to 0$ transitions both sliced off (since if a path has a $0 \to 1$ transition it must have a $1 \to x$ transition).

Two facts from Example 2 are important to remember. First, the structure of the polytope is related more to the topology of the tree than the size of the tree. Second, there is a distinction between even and odd length paths. We call a binary tree *completely odd* if the tree has all leaves at an odd distance from the root. For example, the tree  is completely odd.

THEOREM 1. *Let T be a completely odd binary tree with more than three nodes. The associated polytope P_T always has the same combinatorial type with 8 vertices and 6 facets (see Figure 2).*

PROOF. First, we derive six inequalities that are satisfied by any binary tree, deriving a "universal" polytope for binary trees. Then we show that a completely odd binary

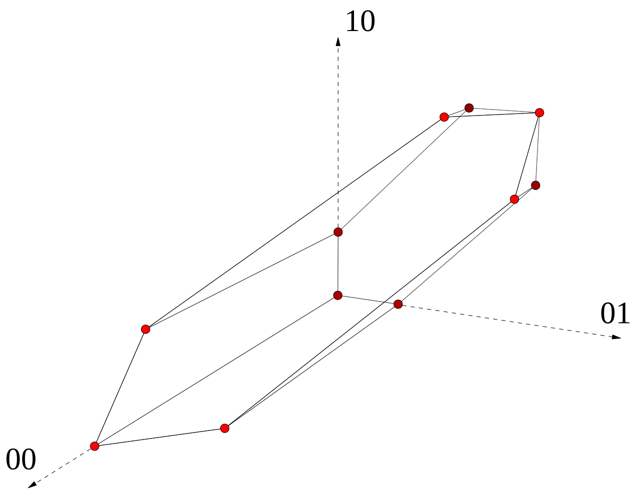


Figure 1: P_T for T a path with 7 nodes, after projecting onto the first three coordinates (b_{00}, b_{01}, b_{10}).

tree has labelings that give us all vertices of the "universal" polytope.

Thinking of the polytope space as the log space of the parameters a_{ij}, we write $\mathbb{R}^4$ with coordinates $b_{00}, b_{01}, b_{10}, b_{11}$. Since P_T lies in $n-1$ times the unit simplex in $\mathbb{R}^4$, we have $b_{00}+b_{01}+b_{10}+b_{11} = n-1$ and the 4 inequalities $b_{ij} \geq 0$. We claim that any binary tree T satisfies two additional inequalities

$$\frac{b_{00}-b_{01}}{2} + b_{10} \leq \frac{n+1}{2}, \tag{2}$$

$$\frac{b_{11}-b_{10}}{2} + b_{01} \leq \frac{n+1}{2}. \tag{3}$$

We prove (2), the second inequality follows by interchanging 1 and 0.

Fix a labeling of the binary tree. We claim that the left hand side of (2) counts the number of zeros that are "created" while moving down the tree, that is, it counts the number of leaves that are zero minus one if the root is labeled zero. Pick a non-leaf of the tree which is labeled "0". It has two children. If both are "0", then this node contributes 2 to $b_{00} - b_{10}$. If both are "1", then this node contributes -2 to $b_{00} - b_{10}$. If one is "0" and one is "1", then the node doesn't contribute. We think of a "0" node with two "0" children as having created a new zero and a "0" node with two "1" children as having deleted a zero. Therefore we see that the term $(b_{00} - b_{10})/2$ counts the number of zeros created as children of "0" nodes. Similarly, if a non-leaf is labeled "1", then its contribution to b_{10} counts the number of new zeros in the children.

Since there are $\frac{n+1}{2}$ leaves in a binary tree, there can be at most $\frac{n+1}{2}$ zeros created, so (2) holds. Notice that the labelings that lie on this facet are exactly those with a one at the root and all zeros at the leaves.

These six inequalities and the equality $b_{00}+b_{01}+b_{10}+b_{11} = n-1$ define a three dimensional polytope in $\mathbb{R}^4$. We compute

Number of nodes	Number of binary trees	Min vertices	Max vertices	Ave vertices
3	1	4	4	4
5	1	7	7	7
7	2	8	10	9
9	3	8	13	11.33
11	6	10	14	11.66
13	11	11	13	11.91
15	23	8	16	14.35
17	46	12	17	13.82
19	98	10	20	14.65
21	207	8	19	14.8
23	451	10	20	15.6

Table 3: Minimum, maximum and average number of vertices of P_T over all binary trees with at most 23 nodes

that there are eight vertices of this polytope:

$$(n-1,0,0,0), \quad (n-3,0,2,0)$$
$$\left(\frac{n-3}{2}, \frac{n+1}{2}, 0, 0\right), \quad \left(0, \frac{2n}{3}, \frac{n-3}{3}, 0\right)$$
$$\left(0, \frac{n-3}{3}, \frac{2n}{3}, 0\right), \quad \left(0, 0, \frac{n+1}{2}, \frac{n-3}{2}\right)$$
$$(0,2,0,n-3), \quad (0,0,0,n-1)$$

Six of these vertices occur in any binary tree: a tree with all zeros gives the $(n-1,0,0,0)$ vertex, a tree with a one at the root and zeros elsewhere gives $(n-3,0,2,0)$, and a tree with ones at the leaves and zeros elsewhere gives $(\frac{n-3}{2}, \frac{n+1}{2}, 0, 0)$. Interchanging 1 and 0 gives three more vertices. However, the remaining two vertices aren't obtained by all binary trees.

The vertex $(0, \frac{n-3}{3}, \frac{2n}{3}, 0)$ lies on the facet defined by (2), so we know it must have a one at the root, all zeros at the leaves, and the labels must alternate going down the tree since there are no zero to zero or one to one transitions. This means that this vertex is representable by a labeled tree if and only if the tree has all leaves at an odd depth from the root. Notice that this implies that n must be divisible by 3 for the tree to be completely odd. Finally, if $n > 3$ is odd and divisible by 3, then $n \geq 9$ and one checks that the eight vertices are distinct.

See Figure 2 for a picture of the polytope and a Schlegel diagram with descriptions of the labelings on the facets and at the vertices. $\square$

In the case where T is binary but not completely odd, the polytope shares 6 vertices with this universal polytope, but the remaining 2 vertices are either not integral or not realizable. However, the polytope still shares much of the boundary with the universal polytope, so it is perhaps realistic to expect that the polytope for a general binary tree behaves well. Table 3 shows data from computations for all binary trees with at most 23 nodes. The maximum number of vertices of P_T appears to grows very slowly with the size of the tree.

Although binary trees seem to generally have polytopes with

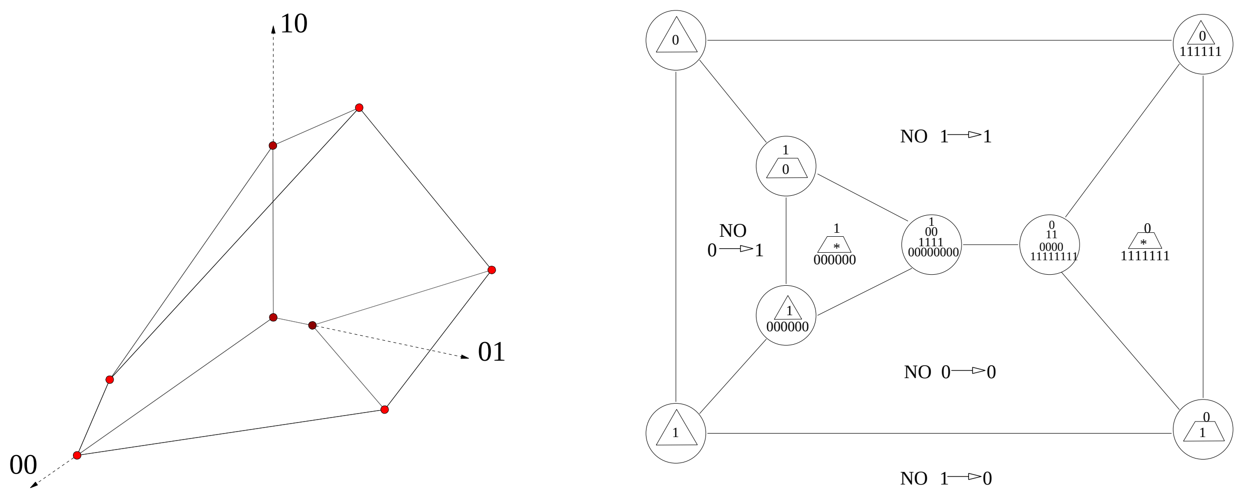


Figure 2: The polytope of the completely odd binary tree and a Schlegel diagram of this polytope with facets and vertices labeled.

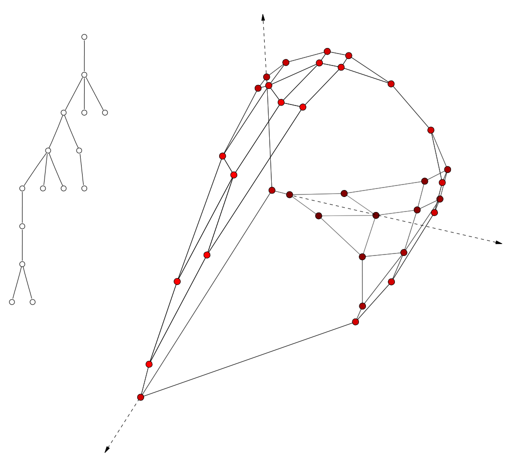


Figure 3: A tree T with 15 nodes for which P_T has 34 vertices, 58 edges, and 26 faces.

Number of nodes	Number of trees	Min vertices	Max vertices	Ave vertices
3	2	4	7	5.5
4	4	4	8	7
5	9	4	11	8
6	20	4	14	9.7
7	48	4	15	10.75
8	115	4	20	12.59
9	286	4	21	13.67
10	719	4	22	15.42
11	1842	4	25	16.60
12	4766	4	28	18.3
13	12486	4	31	19.5
14	32973	4	32	19.75
15	87811	4	34	22.6

Table 4: Minimum, maximum and average number of vertices of P_T over all trees with at most 15 nodes

few vertices, arbitrary trees are not so nice. For example, Figure 3 shows a tree with 15 nodes that has a polytope with 34 vertices.

Table 4 shows data for all trees on at most 15 nodes. It appears that the maximum number of vertices for the polytope of an arbitrary tree of size n grows approximately as $2n$. Notice that the tree with all leaves at depth 1 has P_T a tetrahedron, giving the unique minimum number, 4, of vertices for all trees.

Conjecture 4. *There is a bound on the number of vertices of P_T, where T ranges over all binary trees. However, for an arbitrary tree, the number of vertices of P_T is unbounded.*

To extend these computations, a better algorithm for computing P_T needs to be developed. The naive algorithm for computing P_T involves a loop of size 2^n, elimination of duplicates points, and a convex hull computation. This algorithm can certainly be improved, but it is not known whether there is a polynomial time algorithm for constructing the polytope given a tree. Is there a fast algorithm that, given a tree T and a point of $\mathbb{R}^4$, outputs whether that point arises from a labeling of T? If so, then P_T could be constructed by testing the $O(n^{1.5})$ points inside $n-1$ times the unit simplex.

4. ACKNOWLEDGMENTS

The author is grateful for useful conversations with Seth Sullivant, Lior Pachter, and Bernd Sturmfels and for much help from Raymond Hemmecke in fine tuning `4ti2` for these computations. The polytopal computations were done with `polymake`. The author was supported by a National Defense Science and Engineering Graduate Fellowship.

5. REFERENCES

[1] E. Allman and J. Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical Biosciences*, pages 1–33, 2003.

[2] J. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states.

Journal of Classification, 4:57–71, 1987.

[3] S. Collart, M. Kalkbrener, and D. Mall. Converting bases with the Gröbner walk. *J. Symbolic Comp.*, 24(3):465–469, 1997.

[4] W. Fulton. *Introduction to toric varieties*, volume 131 of *Annals of Mathematics Studies*. Princeton University Press, Princeton, NJ, 1993. The William H. Roever Lectures in Geometry.

[5] R. Hemmecke and R. Hemmecke. 4ti2 version 1.1—computation of Hilbert bases, Graver bases, toric Gröbner bases, and more. Available at `www.4ti2.de`, Sept. 2003.

[6] A. N. Jensen. CaTS, a software system for toric state polytopes. Available at `http://www.soopadoopa.dk/anders/cats/cats.html`

[7] E. Kuo. Viterbi sequences and polytopes, in preparation.

[8] L. Pachter and B. Sturmfels. The geometry of statistical models for biological sequences, eprint: `arXiv q-bio.QM/0311009`

[9] M. Steel and Y. Fu. Classifying and counting linear phylogenetic invariants for the Jukes-Cantor model. *Journal of computational biology*, 2(1):39–47, 1995.

[10] M. Steel, L. Székely, P. Erdös, and P. Waddell. A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. *New Zealand Journal of Botany*, 31:289–296, 1993.

[11] B. Sturmfels. *Gröbner bases and convex polytopes*, volume 8 of *University Lecture Series*. American Mathematical Society, Providence, RI, 1996.

[12] B. Sturmfels. Equations defining toric varieties. In *Algebraic geometry—Santa Cruz 1995*, volume 62 of *Proc. Sympos. Pure Math.*, pages 437–449. Amer. Math. Soc., Providence, RI, 1997.